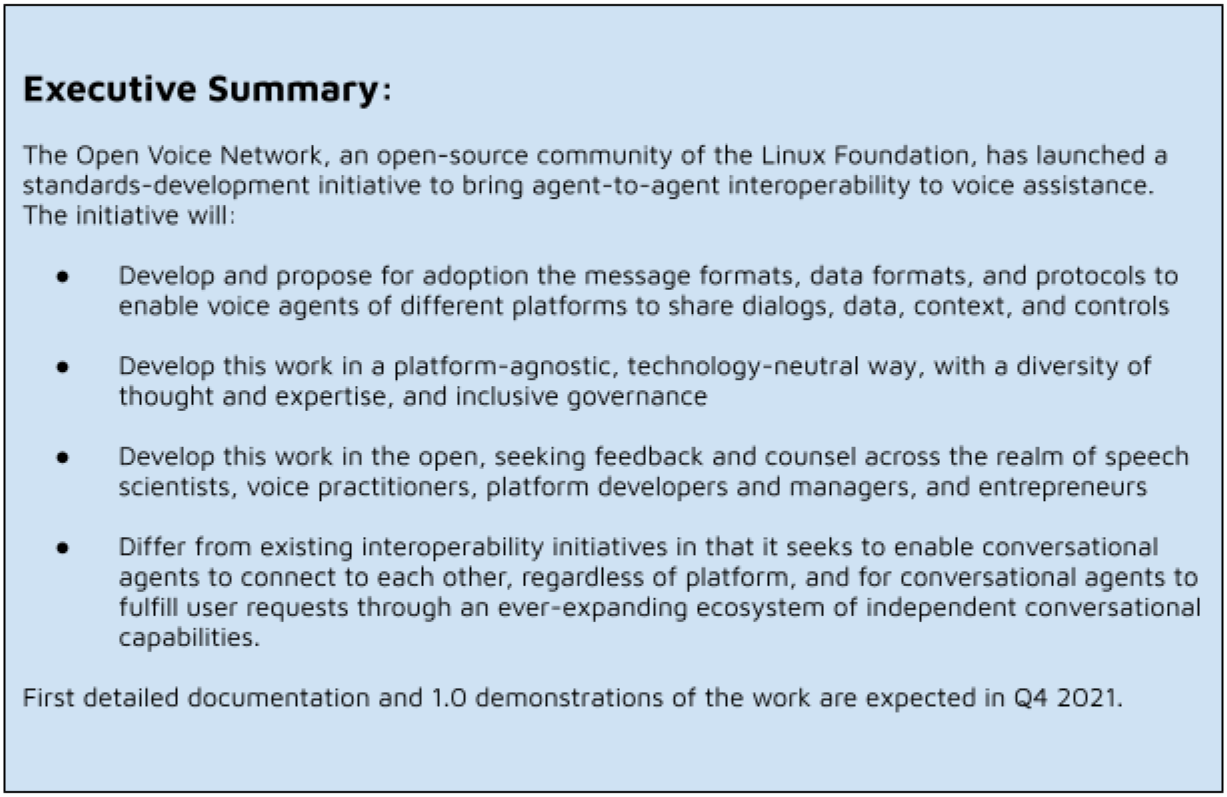
OPEN VOICE NETWORK
Voice for everyone

Part of an introductory overview series of current research, hypotheses, and proposals on responsibly integrating voice technology into society by members of the Open Voice Network, an open-source community of The Linux Foundation

VOICE AGENT INTEROPERABILITY

This overview of the Open Voice Network's approach to voice agent interoperability is intended to provide a lay-level introduction for prospective participants in and sponsors of this important work.

Documentation of the Open Voice Network (OVON) Interoperability Initiative will be stored in the Open Voice Network's GitHub repository, which can be accessed here: Open Voice Network · GitHub.



Executive Summary:

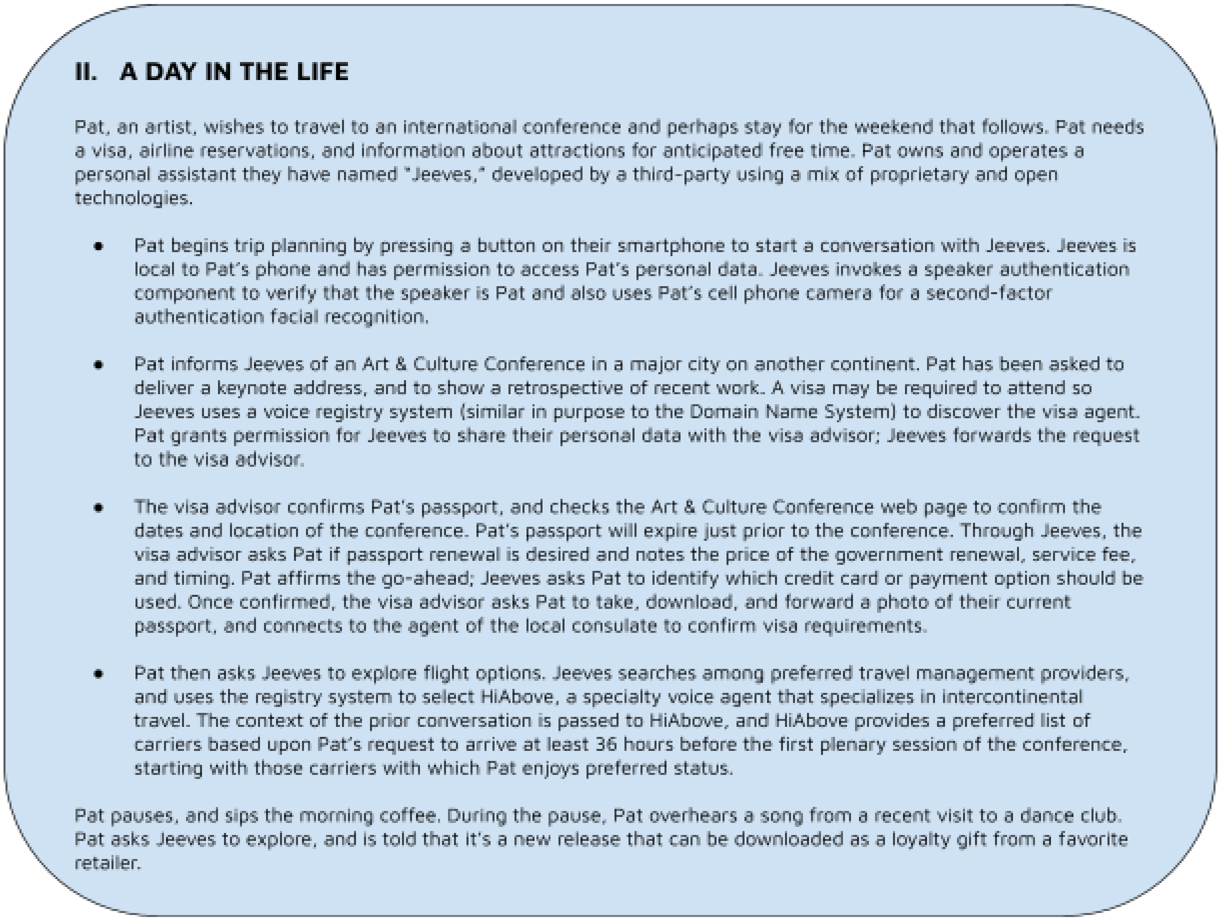
The Open Voice Network, an open-source community of the Linux Foundation, has launched a standards-development initiative to bring agent-to-agent interoperability to voice assistance. The initiative will:

- Develop and propose for adoption the message formats, data formats, and protocols to enable voice agents of different platforms to share dialogs, data, context, and controls
- Develop this work in a platform-agnostic, technology-neutral way, with a diversity of thought and expertise, and inclusive governance
- Develop this work in the open, seeking feedback and counsel across the realm of speech scientists, voice practitioners, platform developers and managers, and entrepreneurs
- Differ from existing interoperability initiatives in that it seeks to enable conversational agents to connect to each other, regardless of platform, and for conversational agents to fulfill user requests through an ever-expanding ecosystem of independent conversational capabilities.

First detailed documentation and 1.0 demonstrations of the work are expected in Q4 2021.

I. Working Definition of Interoperability

The ability for voice agents of different platforms or technological parentage to share dialogs, data, context, and controls.



II. A DAY IN THE LIFE

Pat, an artist, wishes to travel to an international conference and perhaps stay for the weekend that follows. Pat needs a visa, airline reservations, and information about attractions for anticipated free time. Pat owns and operates a personal assistant they have named "Jeeves," developed by a third-party using a mix of proprietary and open technologies.

- Pat begins trip planning by pressing a button on their smartphone to start a conversation with Jeeves. Jeeves is local to Pat's phone and has permission to access Pat's personal data. Jeeves invokes a speaker authentication component to verify that the speaker is Pat and also uses Pat's cell phone camera for a second-factor authentication facial recognition.

- Pat informs Jeeves of an Art & Culture Conference in a major city on another continent. Pat has been asked to deliver a keynote address, and to show a retrospective of recent work. A visa may be required to attend so Jeeves uses a voice registry system (similar in purpose to the Domain Name System) to discover the visa agent. Pat grants permission for Jeeves to share their personal data with the visa advisor; Jeeves forwards the request to the visa advisor.

- The visa advisor confirms Pat's passport, and checks the Art & Culture Conference web page to confirm the dates and location of the conference. Pat's passport will expire just prior to the conference. Through Jeeves, the visa advisor asks Pat if passport renewal is desired and notes the price of the government renewal, service fee, and timing. Pat affirms the go-ahead; Jeeves asks Pat to identify which credit card or payment option should be used. Once confirmed, the visa advisor asks Pat to take, download, and forward a photo of their current passport, and connects to the agent of the local consulate to confirm visa requirements.

- Pat then asks Jeeves to explore flight options. Jeeves searches among preferred travel management providers, and uses the registry system to select HiAbove, a specialty voice agent that specializes in intercontinental travel. The context of the prior conversation is passed to HiAbove, and HiAbove provides a preferred list of carriers based upon Pat's request to arrive at least 36 hours before the first plenary session of the conference, starting with those carriers with which Pat enjoys preferred status.

Pat pauses, and sips the morning coffee. During the pause, Pat overhears a song from a recent visit to a dance club. Pat asks Jeeves to explore, and is told that it's a new release that can be downloaded as a loyalty gift from a favorite retailer.

III. The Voice-Expanded Web: The World for Which We Build

Voice assistance—the artificial intelligence-supported technologies that enable natural language communication between humans and devices of all types, from smartphones to automobiles—is becoming a primary interface to the digital world. Natural language inputs are joining the ranks of, and in many cases replacing, the classic data inputs: typing, taps, and swipes. Natural language outputs also are delivering information directly and in concert with screen-based visuals, a feature of Universal Design, which encompasses accessibility and inclusion.

This transition in consumer behavior is leading us into a voice-expanded World-Wide Web, one that combines artificial intelligence-powered applications such as natural language understanding, natural language generation, and speech recognition, and ambient computing to connect users to billions of destinations. This will include not just voice-enabled websites, but the voice-enabled artificial intelligences of automobiles, smart phones, IOT sensors and smart systems, media properties, and virtual personages and places.

This transition is now underway: A May 2021 study of consumer behavior across the United Kingdom, Germany, and the United States[1] showed that more than 95 percent of adults in all three nations were aware of voice assistance, and roughly one-third of all adults used it daily or multiple times a day. Device usage has not been limited to smart speakers; a February 2021 study of U.S. voice assistant usage within the prior year revealed that 211.5 million persons had accessed voice assistance through a smartphone; 123.8 million within a smart home; 94.7 million via wearables (such as Apple's AirPods); and 88.5 million through smart speakers.[2]

As this transition unfolds, we expect to hear of more users searching the web by voice, and jumping to-and-from voice-enabled websites. Users will also have the option to use voice to:

- review business data;
- adjust ambient lighting;
- guide transportation systems;
- explore virtual worlds with fictional characters and synthetic voices; and,
- manage their personal assistant, and share personal data through voice-based conversation and consent.

Should this transition be supported and accelerated by broadly adopted open technology standards, significant growth can be predicted in both the size and breadth of the market.[3] Value will be realized through the ability to connect, to create layers of ever-more valuable connection, and the acquisition and deep analysis of data. Agent-to-agent interoperability addresses not just an issue of technology, but the economic question that looms before the voice-expanded World Wide Web: who will reap the value?

[1] Vixen Labs and Pragmatic Digital, *Voice Consumer Index*, as executed by Delineate, London. May 2021
[2] Voicebot.ai, February 24, 2021
[3] Padilla, Davies, and Boutin, *Economic Impact of Technology Standards*, Compass Lexecon, September 2017.

IV. Open Voice Network 1.0 Hypotheses for Agent-to-Agent Interoperability

Standardized message formats, data formats, and protocols will enable agent-to-agent interoperability. The establishment of standardized messaging format, data formats, and protocols will be accelerated through the adoption and adaptation of existing formats, protocols, and languages. For the Open Voice Network, these will include:

- existing standards, protocols, tools, and technologies such as MRCP, WebRTC, W3C SSML, Emma, and others; and,

- existing research from such leading voice assistance centers as the Stanford Open Voice Assistant Laboratory (OVAL) and the W3C Voice Interaction Community.

The current (Q3 2021) economics of the current voice assistance industry may create resistance to the concept of agent-to-agent interoperability. However, the economic opportunities within a voice-expanded World Wide Web will drive—and ultimately pull— the adoption of standardized messaging formats, protocols, and languages.

V. Vocabulary: Working Definitions

- **Conversational access point**: the means of carrying a signal to a conversational platform. Users initiate and conduct conversations through **access points**. An **access point** can be physical or virtual. Current examples of conversational **access points** include smartphones, smart speakers, personal computers, wearables, or microphone-enabled smart home hubs.
  - *In the interoperable world, **conversational access points** will enable connection to any conversational platform and agent.*

- **Conversational platform**: the combination of components that enables the operation and management of one or more conversational agents. These components may include artificial intelligence-powered natural language understanding, natural language generation, and dialog management. Current examples of conversational platform providers are Amazon, Google, Xiaomi, Nuance, SoundHound, Rasa, Cerence, Mycroft, and many others[4].
  - *In the interoperable world, creators of conversational agents will be able to choose from multiple **conversational platform** providers.*

[4] Names and brands may be the property of other organizations.

- **Conversational agent:** perceived by users to be a single conversation actor. It uses the infrastructure of the conversational platform, and uses one or more conversational capabilities to hold conversations with the user, with continuity of knowledge and persona and a name by which it can be addressed.
  - *In the interoperable world, there will be both general-purpose and specific-purpose **conversational agents**. Dialogs, data, context and controls will be passed between **agents** according to standardized protocols.*

- **Conversational sub-agent:** a conversational agent that *can only be invoked through another agent*. It performs a delegated task on behalf of another conversational agent. Sub-agents have a name and their own discourse context. **Sub-agents** are sometimes commonly referred to as "skills" or "actions."
  - *In the interoperable world, **sub-agents** will likely emerge as fully realized agents.*

- **Conversational capability:** provides specific dialog functions that inform an agent. **Capabilities** do not have a defined name or voice, nor the continuity of discourse context. A **capability** may be as limited as a response from an IOT sensor or an information fragment in support of an agent dialog; it may be as expansive as a partial or full dialog on behalf of an agent.
  - *In the interoperable world, **conversational capabilities** will exist independently from agents.*

V. Why This and Why Now

The challenge of the current state for individual and enterprise/organizational users is that the value of voice assistance today is bound by the many limitations of proprietary conversational platforms. In these walled technology gardens:

- Users are unable to communicate directly with and through other agents, or access capabilities developed on other platforms.
- Users—individual and organizational users alike— are expected to share all data with the platform provider, which could include conversational, tonal, biometric, and biomarker data.
- Users must surrender credentials to access in-platform capabilities.
- Sub-agent and capability creators and voice innovators are limited to the technology dictates of the proprietary platforms.

The lack of interoperability between market-leading, proprietary conversational agents requires that organizational users of voice must develop and maintain multiple, proprietary systems, or invest in agent translation and management software. This is among the many reasons a recent European Commission sector inquiry[5] found that proprietary conversational platforms and their agents limit the market potential and current growth of home-based Internet of Things (IOT) systems.

The benefits of agent-to-agent interoperability:

- **For individual users of voice assistance:** standards-based agent-to-agent interoperability will enable individuals to locate and connect with any standards-based voice agent or voice-enabled, standards-based IOT destination.

- **For enterprise and organizational owners of voice assistance**: standards-based agent-to-agent interoperability will enable enterprises and organizations to:
    - Build, maintain, operate, and innovate through a single voice agent and its branded persona, without the cost of multi-platform management or multi-platform translation;
    - Reach all constituents, regardless of originating access point and platform;
    - Pursue the economic value of voice on a level playing field, free of proprietary rules and the threat of disintermediation.

- **For developers and designers of voice assistance:** standards-based agent-to-agent interoperability will enable developers and designers to:
    - create un-branded, independent capabilities for sale, re-sale, and customization to developers and managers of voice agents;
    - reduce time spent on basic functionality;
    - innovate from a standards-based foundation to create new, valuable capabilities;
    - participate in a growing ecosystem of training, tool kits, reusable and customizable capabilities, and development environments.

We believe the future of voice assistance is a voice-expanded World Wide Web, made of:

- general purpose, proprietary platform-based conversational agents
- brand-, title-, character- enterprise- and organization-specific conversational agents
- sensors, aggregators, and AI's within smart, IOT environments
- capabilities with generic, reusable functions that can be used to accelerate the development of conversational agents.

[5] *Commission Staff Working Document: Preliminary Report—Sector Inquiry into Consumer Internet of Things*, European Commission, Brussels, September 6, 2021

VI. An Interoperable Voice Agent Should…

- Reuse conversational capabilities from other developers, including those running on other conversational platforms
- Share dialog, data, context, and controls with other voice agents, regardless of platform or technological parentage
- Be addressable directly through an explicit user request
- Make itself available to meet a user request on behalf of another trusted voice agent
- Support security and privacy by design.

VII. The Open Voice Network Interoperability Initiative: From Talk to Action

To develop the standard message formats, protocols, and data formats necessary for agent-to-agent interoperability, the Open Voice Network will cut the issue into layers. Presently, we envision six layers, termed “enhancements,” that operate atop a standardized communication infrastructure.

This infrastructure will include:

- a set of well-defined messages for starting, stopping, suspending, and resuming voice agents;
- protocols consisting of sequences of messages that implement the enhancement; and,
- a set of self-defined data formats for transferring data among agents.

The six enhancements are as follows:

1. **Remote invocation and location of a second voice agent**, regardless of platform.

2. **Connection to a specific agent through an explicit request**. Such a request could contain the registered name of a voice agent. The Open Voice Network envisions a “Voice Registration System” (VRS) that allows an owner of voice agents to uniquely register the name of the agent, and enables a direct connection to that agent when that agent is explicitly requested by the user.

3. **Switching between voice agents**. To fulfill user needs, a voice agent may need to switch the user to a second voice agent—and then back again. For instance, an experienced

traveler may wish to switch back-and-forth between an airline-specific agent, a hotel chain-specific agent, and the agents of restaurants.

4. **Processing implicit requests**. We define "implicit" as a request that searches across voice agent services rather than seeking specific, named voice agent destination. This will rely upon an arbitration service to select (or recommend) a voice agent from many.

5. **Sharing data and context between voice agents.** This enhancement advances #3, above, and enables agents to manage the sharing of data, context, and controls. It also will enable agents to call upon the universe of conversational capabilities, regardless of platform parentage.

6. **Generating and operating a private agent persona.** A persona refers to the voice and characteristics presented by a conversational agent. Such agents have a distinctive sound and ambiance that leads to branding, the promotion of a particular enterprise by means of advertising, and distinctive design. This enhancement will enable an enterprise voice agent to fully express a specific brand tone, and enable capabilities to adapt to or adopt the persona of the enterprise agent.

Other possible enhancements include:

- discovery—locating an agent;
- planning—developing a sequence of voice agents and capabilities that perform a complex task; and,
- identification and authentication of both the user and desired conversational agent, to allow a secure connection.

These and other potential enhancements are currently outside of the scope of the Open Voice Network Voice Agent Interoperability initiative.

VIII. How You Can Accelerate the Open Voice Network Agent-to-Agent Initiative

To accelerate the Open Voice Network agent-to-agent interoperability initiative, please consider helping to shape the future as you:

- **Invest your expertise in the Open Voice Network interoperability Initiative**:
  - Participate in weekly debates, discussions, and reviews. Meetings are virtual, recorded, no more than 60 minutes, and, in keeping with OVON culture, they begin and end on time.
  - Review and contribute to the work as your schedule permits. Documents (including this one) can be found in the Open Voice Network GitHub repository. Should you see a reason for revision, simply open a pull request. And, you can keep up with the weekly debates and discussions by reviewing the meeting recordings. They're stored in the Open Voice Network Google Workspace, accessible by members and by requesting permission.

- **Raise your voice on behalf of the Open Voice Network interoperability initiative**:
  - Join the Open Voice Network Ambassador corps. OVON Ambassadors participate regularly in public forums, whether through social media, social audio, or industry events. You'll have the support of the OVON's Outreach Team.

- **Have your firm provide financial sponsorship for this important work**:
  - There are OVON sponsorship opportunities for every size of organization.
  - Sponsor monies allow the OVON to hire specialists to accelerate research, write code, and develop prototypes and demos, as well as enable the day-to-day operation and outreach of this global, open-source community.

IX. Frequently Asked Questions

Q: Why should I participate in the Open Voice Network Voice Agent Interoperability Initiative?

A: More than 170 speech scientists, voice developers, and voice practitioners across 13 nations now contribute regularly to the work of the Open Voice Network. Broadly stated, these individuals believe that:

- Voice assistance will reach its optimum value worldwide for developers, enterprises, and voice platforms only if it is open, standards-based, and worthy of user trust.
- Standards are of the greatest value when they are created through neutral, communal, and diverse research, debate, and development.
- Standards drive and expand technology innovation and economic opportunity.

Q: How does the Open Voice Network Voice Agent Interoperability Initiative (VAII) differ from other voice interoperability initiatives?

A: The Open Voice Network applauds—and welcomes collaboration with—all initiatives that seek to expand the value of voice assistance through interoperability. In the simplest terms, the Open Voice Network seeks a future for voice that resembles today's internet: an open, standards-based, voice-expanded and voice-enabled world-wide web, where billions of specific-purpose, unique-persona agents connect and share dialogs, data, context, and controls, and independent conversational capabilities abound.

To achieve this vision, we propose a software-centric, agent-to-agent approach to voice interoperability standardization, one pursued through communal debate and development that is platform-agnostic and technology-neutral.

There are now initiatives in the market that define interoperability in what might be described as a hardware-centric approach. This approach would allow conversational agents of different parentage to reside on a single device; users would toggle back and forth between agents, stopping and starting dialogs as needed. This is clearly useful in certain situations; indeed, the Open Voice Network expects to explore and develop standards for this functionality (see enhancement #3, above.)

Q: Will the Open Voice Network develop an open conversational platform?

A: No. The Open Voice Network will not develop an open, standards-based platform.

We seek to develop standards that will enable conversational agents of different platforms (open, proprietary, hybrid) to share dialogs, data, context, and controls.

Q: Will developers be able to insert Open Voice Network enhancements directly into their products?

A: No. Enhancements will likely need to be refined and tailored to be effective in each vendor product as they use different platforms, languages, and implementation strategies. The true value of Open Voice Network interoperability work will most likely be found in the standard message formats, protocols, and data formats, and the algorithms used to implement each enhancement.

IX. About the Open Voice Network (www.openvoicenetwork.org)

Vision

The Open Voice Network (OVON) seeks to make voice technology worthy of user trust—a task of critical importance as voice emerges as a primary, multi-device portal to the digital and IOT worlds, and as independent, specialist voice assistants take their place next to general purpose platforms.

Mission

The Open Voice Network will achieve its vision through the communal development and adoption of industry standards and usage guidelines, industry education and advocacy initiatives, and the development and documentation of voice-centric value propositions.

Principles

The Open Voice Network is guided by four values. It seeks a world of voice technology that:

1. is worthy of user trust;
2. enables user, ecosystem, and architectural choice;
3. is inclusive and accessible; and
4. is open in software and hardware, with standards serving as a foundation for commercial differentiation.

History and Affiliations

The Open Voice Network was created on June 1, 2020 as an open-source community of the Linux Foundation. Its origins can be traced to research by the Massachusetts Institute of Technology

(MIT) Auto-ID Laboratory, Capgemini Consulting, and the Intel Corporation, and corporate seed funding provided in 2019.

Governance

The Open Voice Network is an independently funded and governed non-profit industry association that operates as an open source community of The Linux Foundation.

- Financial support for OVON comes from sponsoring enterprises and organizations. Governance and strategic direction is provided by a Steering Committee of senior executives of sponsor organizations.
- The Executive Director, who reports to the Steering Committee, is responsible for day-to-day management and the chairs of the OVON standing committees and moderators of OVON communities.

- As an open source community of the Linux Foundation, OVON enjoys access to the expertise and shared legal, operational, and marketing services of a world leader in the creation of open-source projects and ecosystems.

X. About The Linux Foundation

The Linux Foundation is dedicated to building sustainable ecosystems around open-source projects to accelerate technology development and industry adoption. Founded in 2000, the Linux Foundation provides unparalleled support for open-source communities through financial and intellectual resources, infrastructure, services, events, and training. Working together, the Linux Foundation and its projects form the most ambitious and successful investment in the creation of shared technology.

For more information, please visit https://www.linuxfoundation.org/.